To impute or not to impute categorical data: applications of multiple correspondence analysis

Johané Nienkemper-Swanepoel

Stellenbosch University, South Africa nienkemperj@sun.ac.za

Niël J le Roux

Stellenbosch University, South Africa

Sugnet Lubbe

Stellenbosch University, South Africa

Keywords

Biplots, Generalised Procrustes analysis, Missing categorical data, Multiple imputation, Subset multiple correspondence analysis

This presentation will address the application of multiple correspondence analysis (MCA) in the visualisation of multivariate categorical data containing missing observations. A multiple imputation algorithm that relies on MCA is applied to obtain completed data sets. The completed data sets are then analysed using MCA biplots. As the number of multiple visualisations increase with multiple imputation it becomes a cumbersome interpretation task. A combining procedure is developed, referred to as GPAbin, which consists of applying generalised orthogonal Procrustes analysis (GPA) and obtaining a final combined configuration thereafter. The GPAbin procedure results in the unbiased representation of multiple MCA biplots. Another approach is to avoid imputation by recoding the indicator matrix of the data set and creating new category levels to which missing values are assigned. The recoded indicator matrix is used for subset MCA (sMCA) which enables separate visualisations for the complete and missing parts of the data. A comparison between the imputation and non-imputation approaches will be presented highlighting the advantages and disadvantages of both approaches. A simulation study is reported where data sets with various sample sizes, variables and varying number of category levels are simulated from three different distributions. The presentation of the results will focus on the influence of the underlying distribution and characteristics in the simulated data that leads to the success or downfall of the proposed methods. The results from the simulation study aid as a guide when deciding to apply an imputation or non-imputation technique for the analysis of real data.